

BUSTA A

1. Discuti tecniche, problematiche e tematiche relative a:
 - Exploratory Data Analysis, data preprocessing, data cleaning, e dimensionality reduction. Contestualizza a datasets con problematiche di multicollinearità, missing values (valori mancanti), scarso bilanciamento dei dati, distribuzioni caratterizzate da skew, etc.
2. Rispondere in inglese: Discuss techniques, problems, and solutions related to:
 - Systems and libraries for handling large amounts of data. Contextualize to concepts related to relational databases, SQL queries, data lakes, etc. Also discuss these issues in the context of HPC systems.
3. Esercizio di programmazione: usare uno pseudo-linguaggio, o un linguaggio di programmazione a scelta tra Python, R, C/C++, Java, Bash.

Leggere da terminale una sequenza di lettere minuscole. La sequenza è da considerarsi terminata quando viene letto il carattere 0. Tale carattere è da considerarsi un carattere speciale di terminazione, da non considerare nella lista di lettere ricevute.

In input possono trovarsi sequenze con ripetizioni consecutive della stessa lettera. Determinare la lunghezza massima di tali sequenze. Per esempio, per l'input

```
a c a b c c c c 0
```

la lunghezza cercata è 4, per via delle 4 ripetizioni della lettera c.

Determinare e stampare la lista di lettere inserite, in ordine alfabetico, senza ripetizioni. Creare e popolare una struttura dati adeguata per associare ogni lettera diversa inserita con il numero di volte che è stata inserita.

BUSTA B

1. Discuti tecniche, problematiche e tematiche relative a:
 - Unsupervised machine learning. Contestualizza a task di clustering. Si pensi ad esempi come il clustering di cifre scritte a mano del 'MNIST' database che contiene immagini o loro versioni tabellari.
2. Rispondere in inglese: Discuss methodologies and systems for the design of data bases in a high I/O rate environment. Discuss the phases of conceptual modeling, and of logical modeling. Highlight the main goals and tasks of each of the two phases, and how they differ from each other.
3. Esercizio di programmazione: usare uno pseudo-linguaggio, o un linguaggio di programmazione a scelta tra Python, R, C/C++, Java, Bash.
Leggere da terminale una sequenza di parole (sequenze alfanumeriche tipicamente rappresentate con il tipo di dato 'string' nei linguaggi di programmazione). Le parole sono composte solo da lettere minuscole. La sequenza è da considerarsi terminata quando viene letta la parola "end". Tale parola è da considerarsi un carattere speciale di terminazione, da non considerare nella lista di parole ricevute.
Stampare in output alcune proprietà della lista ricevuta in input: quante parole sono state lette, quante parole diverse sono state lette, il numero medio di caratteri presenti in ogni parola, il numero medio di vocali a, e, i, o, u presenti in ogni parola. Creare e popolare una struttura dati adatta ad associare ad ogni lettera il numero di parole inserite aventi quella lettera come iniziale.

BUSTA C

1. Discuti tecniche, problematiche e tematiche relative a:
 - Supervised machine learning. Contestualizza a task di classificazione. Si pensi ad esempi di dataset come il 'Titanic' in cui, in base ad informazioni sui passeggeri, si deve predire se sarebbero sopravvissuti o meno al naufragio.
2. Rispondere in inglese: Two established ways to make computing resources available are virtual machines and containers. Discuss the two approaches. How do the two approaches differ? Which technologies and products can be used in each case?
3. Esercizio di programmazione: usare uno pseudo-linguaggio, o un linguaggio di programmazione a scelta tra Python, R, C/C++, Java, Bash.

Leggere da terminale una sequenza di lettere minuscole. La sequenza è da considerarsi terminata quando viene letto il carattere 0. Tale carattere è da considerarsi un carattere speciale di terminazione, da non considerare nella lista di lettere ricevute.

Calcolare il numero di 'parole' (sequenze di lettere, senza necessariamente alcun significato) diverse di 3 lettere che si possono ottenere usando le lettere ricevute in input. È preferibile evitare di rispondere a questa domanda usando una semplice enumerazione 'brute force' di tutte le parole ottenibili.

Assumendo di avere a disposizione una struttura dati contenente le parole di senso compiuto, calcolare il numero di parole di senso compiuto di 3 lettere che si possono ottenere usando le lettere ricevute in input.